MLE Exercises

1. In this assessment we are going to try to answer the question: is there a section of the human cytomegalovirus genome in which the rate of palindromes is higher than expected?

Make sure you have the latest version of the dagdata library:

```
library(devtools)
install_github("genomicsclass/dagdata")
```

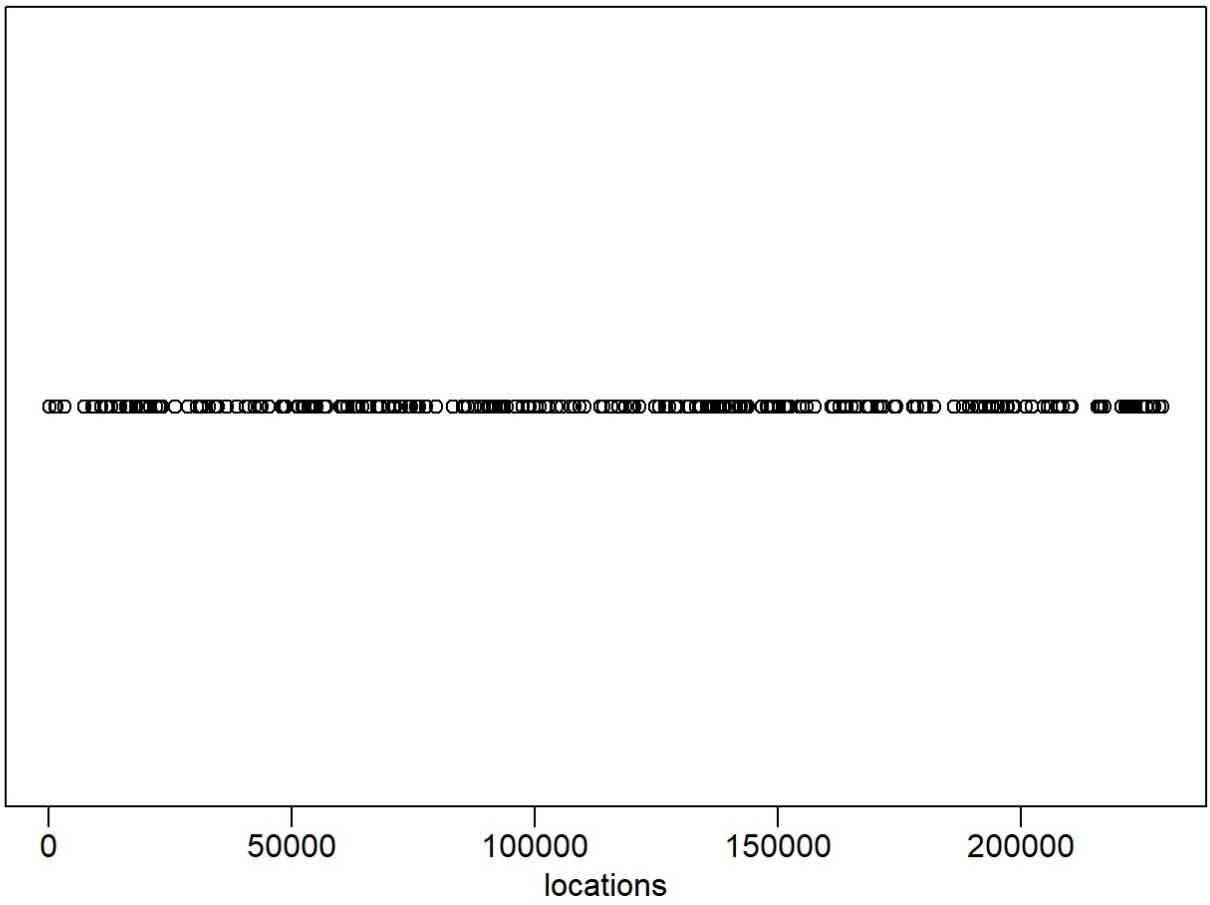
```
## Skipping install of 'dagdata' from a github remote, the SHA1 (b4861304) has not cha
nged since last install.
##   Use `force = TRUE` to force installation
```

```
#and then load the palindrome data from the Human cytomegalovirus genome:

library(dagdata)
data(hcmv)

#These are the locations of palindromes on the genome of this virus:

library(rafalib)
```

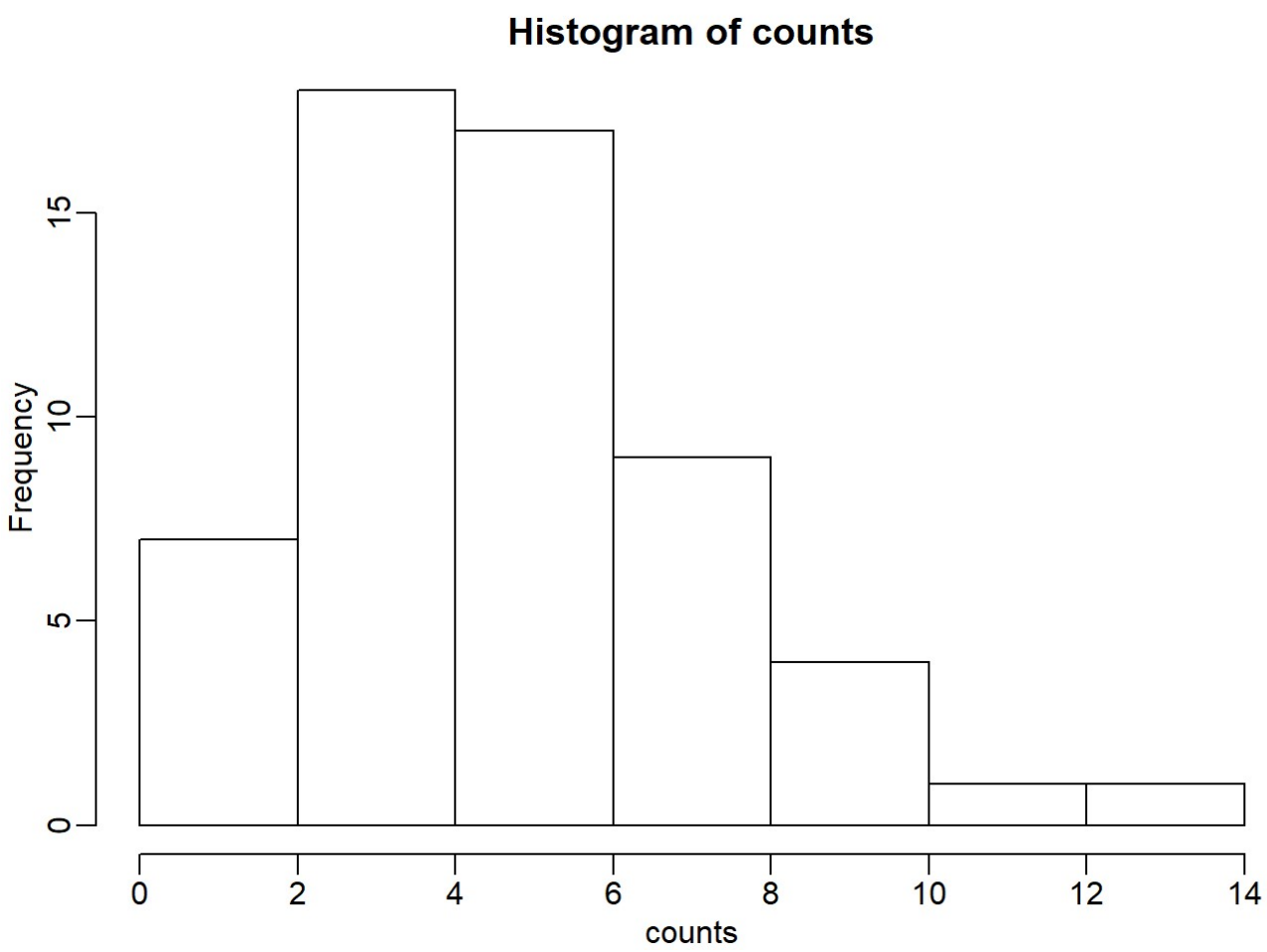
```
##
## Attaching package: 'rafalib'
```

```
## The following object is masked from 'package:devtools':
##
##     install_bioc
```

```
mypar()
plot(locations,rep(1,length(locations)),ylab="",yaxt="n")
```




```
# These palindromes are quite rare,
# is very small. If we break the genome into bins of 4000 basepairs, then we have
#
# not so small and we might be able to use Poisson to model the number of palindromes
in each bin:

breaks=seq(0,4000*round(max(locations)/4000),4000)
tmp=cut(locations,breaks)
counts=as.numeric(table(tmp))

#So if our model is correct counts should follow a Poisson distribution. The distribut
ion seems about right:

hist(counts)
```




```
# So let X1, ..., Xn be the random variables representing counts then Pr(Xi = k) = lam
bda^k/k! exp(- lamdba) and to fully describe this distribution, we need to know lambda
.
#
# To compute the Maximum Likelihood Estimate (MLE) we ask what is the probability of o
bserving our data (which we denote with small caps) for a given :
# L(lambda) = Pr(X1 = x1) * Pr(X2 = x2) * ... * Pr(Xn = xn)
#
# Now we can write it in R. For example for lambda = 4
#
# we have:

probs <- dpois(counts,4)
likelihood <- prod(probs)
likelihood
```

```
## [1] 1.177527e-62
```

```
#Run the code above to note that this is a tiny number. It is usually more convenient
to compute log-likelihoods

logprobs <- dpois(counts,4,log=TRUE)
loglikelihood <- sum(logprobs)
loglikelihood
```

```
## [1] -142.5969
```

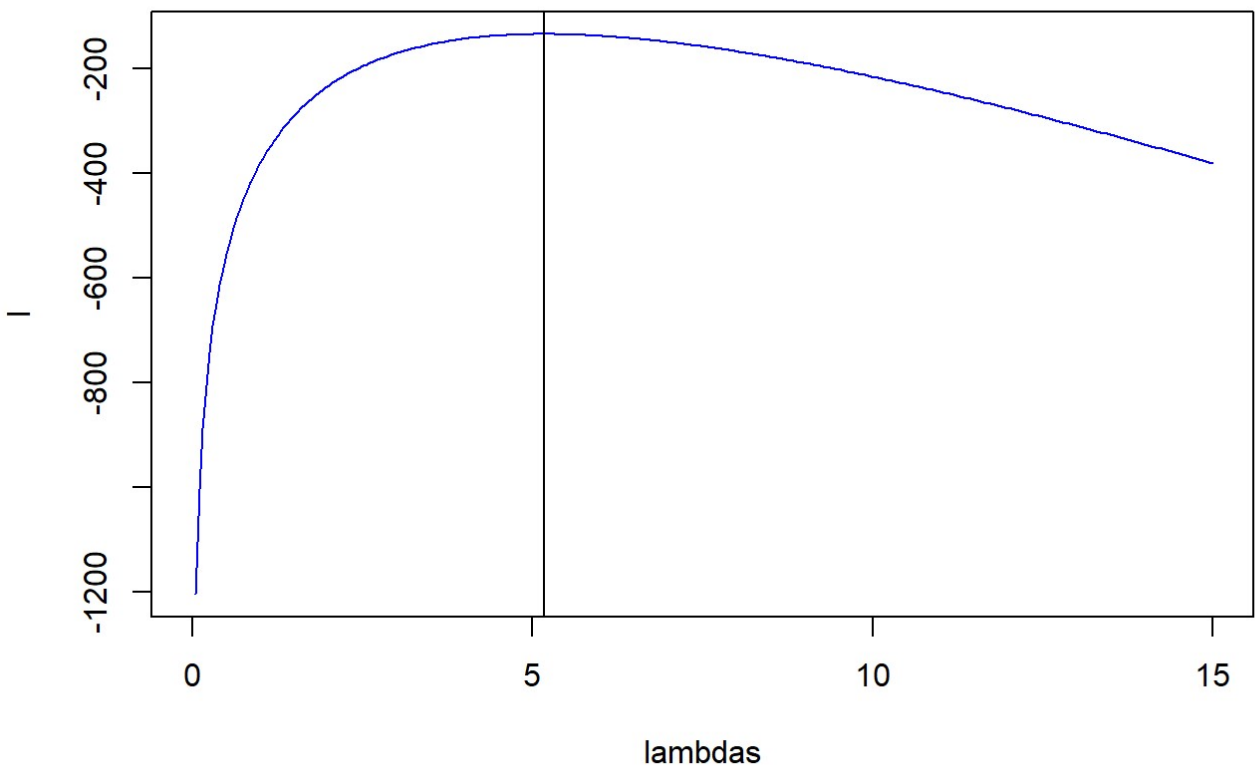
Now write a function that takes lambda and the vector of counts as input, and returns the log-likelihood. Compute this log-likelihood for lambdas = seq(0,15,len=300) and make a plot.

What value of lambdas maximizes the log-likelihood?

```
log_L<-function(lambda,x){
    sum(dpois(x,lambda,log=TRUE))
}
lambdas = seq(0,15,len=300)

l <- sapply(lambdas,function(lambda) log_L(lambda,counts))
plot(lambdas,l,type="l",col="blue")

mle <- lambdas[which.max(l)]
mle
```

```
## [1] 5.167224
```

```
abline(v=mle)
```

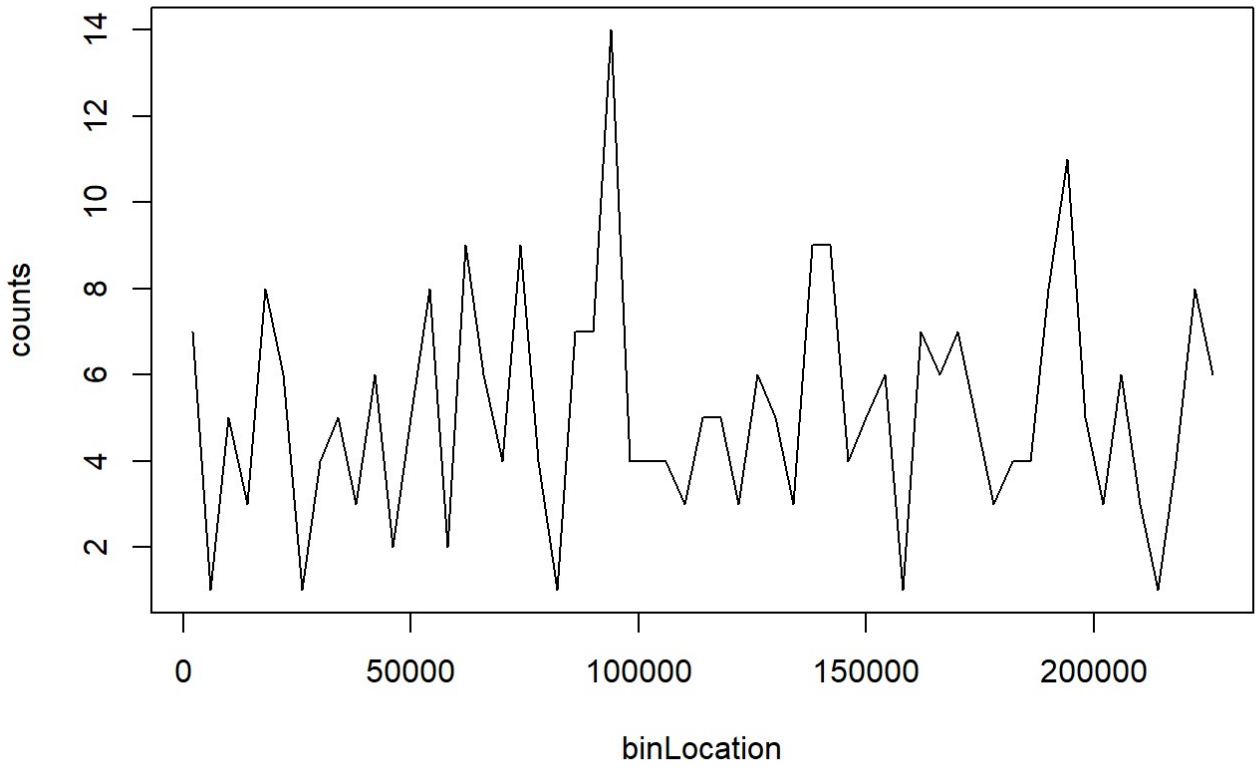


It turns out that, using calculus, we can work out mathematically what lambda

maximizes the likelihood. The average of the counts is the MLE. Note that we obtain a similar number to the answer to Question 4.2.1: mean(counts)

2. The point of collecting this dataset was to try to determine if there is a region of the genome that has higher palindrome rate than expected. We can create a plot and see the counts per location:

```
breaks=seq(0,4000*round(max(locations)/4000),4000)
tmp=cut(locations,breaks)
counts=as.numeric(table(tmp))
binLocation=(breaks[-1]+breaks[-length(breaks)])/2
plot(binLocation,counts,type="l",xlab=)
```



What is the center of the bin with the highest count?

```
binLocation[which.max(counts)]
```

```
## [1] 94000
```

3. For the question above, what is the maximum count?

```
max(counts)
```

```
## [1] 14
```

4. Now that we have identified the location with the largest palindrome count, we want to know if by chance we could see a value this big.

If X is a Poisson random variable with rate lambda = mean(counts[- which.max(counts)])

What is the probability of seeing a count of 14 or more?

```
# 13 p300 in book
lambda <- mean(counts[ - which.max(counts) ])

1 - ppois(13,lambda)
```

```
## [1] 0.00069799
```

5. From the question above, we obtain a p-value smaller than 0.001 for a count of 14. Why is it problematic to report this p-value as strong evidence of a location that is different?

   Poisson in only an approximation. We selected the highest region out of 57 and need to adjust for multiple testing. correct is an estimate, a random variable, and we didn't take into account its variability. We don't know the effect size.

6. Use the Bonferroni correction to determine the p-value cut-off that guarantees a FWER of 0.05. What is this p-value cutoff?

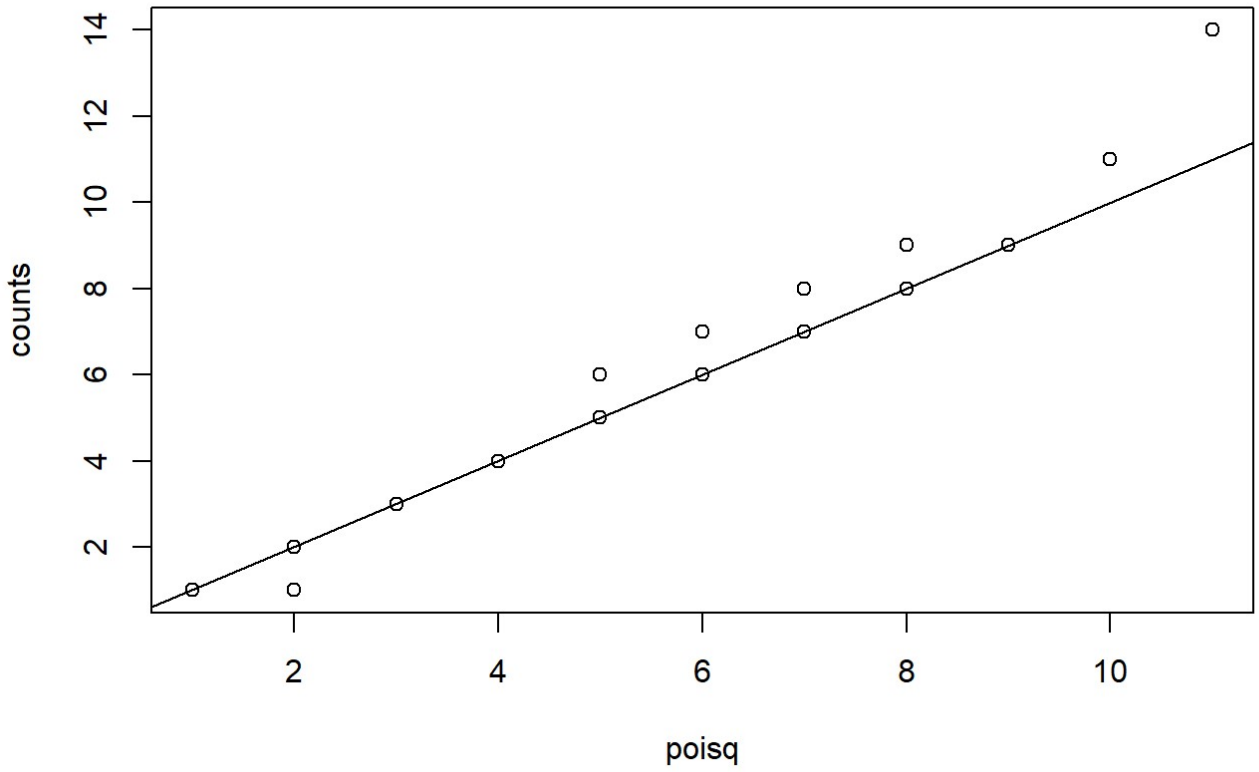
```
alpha <- 0.05
m <- 57
(k <- alpha / m)
```

```
## [1] 0.000877193
```

```
#Note that our observed p-value satisfy the Bonferroni correction.
```

7. Create a qq-plot to see if our Poisson model is a good fit:

```
ps <- (seq(along=counts) - 0.5)/length(counts)
lambda <- mean( counts[ -which.max(counts)])
poisq <- qpois(ps,lambda)
qqplot(poisq,counts)
abline(0,1)
```



How would you characterize this qq-plot

```
Poisson is a terrible approximation.
Poisson is a very good approximation except for one point that we actually think is as
sociated with a region of interest. correct
There are too many 1s in the data.
A normal distribution provides a better approximation.
```